



Sensitivity and calibration disparities of a race-blind machine learning model for periodontitis across racial and ethnic groups

Disparidades de sensibilidad y calibración de un modelo de aprendizaje automático para la periodontitis entre grupos étnicos

Maria Fernanda Escalante Freitez¹  

¹ Florida International University, USA.

Submitted: 06-05-2026 | Revised: 09-07-2026 | Accepted: 10-09-2026 | Published: 14-09-2026

Corresponding Author: Maria Fernanda Escalante Freitez 

ABSTRACT

Introduction: Machine learning models are increasingly proposed for periodontitis prediction, but their fairness across racial and ethnic groups is rarely quantified. **Methods:** A race-blind logistic regression model was trained on 12,630 adults aged 30 years and older from NHANES 2009-2014 to predict periodontitis, and its discrimination, sensitivity, specificity, calibration, and demographic parity were evaluated within each racial and ethnic group. **Results:** The model discriminated periodontitis with an overall area under the curve of 0.656. Sensitivity ranged from 54.1% to 79.2% and specificity from 46.0% to 73.9% across groups, yielding an equal-opportunity gap of 25.1 percentage points and a demographic-parity gap of 29.5 points, while calibration error remained below 12.2% in every group. **Conclusions:** A race-blind periodontitis model produced marked disparities in sensitivity and specificity across racial and ethnic groups through proxy variables, despite comparable calibration.

Keywords: periodontitis; machine learning; fairness; healthcare disparities; calibration.

RESUMEN

Introducción: Los modelos de aprendizaje automático se proponen cada vez más para predecir la periodontitis, pero su equidad entre grupos raciales y étnicos rara vez se cuantifica. **Métodos:** Se entrenó un modelo de regresión logística ciego a la raza con 12 630 adultos de 30 años o más de NHANES 2009-2014 para predecir periodontitis, y se evaluaron su discriminación, sensibilidad, especificidad, calibración y paridad demográfica dentro de cada grupo racial y étnico. **Resultados:** El modelo discriminó la periodontitis con un área bajo la curva global de 0,656. La sensibilidad varió del 54,1 % al 79,2 % y la especificidad del 46,0 % al 73,9 % entre grupos, lo que produjo una brecha de igualdad de oportunidades de 25,1 puntos porcentuales y una brecha de paridad demográfica de 29,5 puntos, mientras que el error de calibración permaneció por debajo del 12,2 % en todos los grupos. **Conclusiones:** Un modelo de periodontitis ciego a la raza produjo marcadas disparidades de sensibilidad y especificidad entre grupos raciales y étnicos a través de variables sustitutivas, pese a una calibración comparable.

Palabras clave: periodontitis; aprendizaje automático; equidad; disparidades en salud; calibración.

INTRODUCTION

Periodontitis is among the most prevalent chronic inflammatory conditions in adults, and its distribution across the United States population is sharply unequal. The CDC/AAP surveillance case definition estimated that nearly half of adults aged 30 years and older had some form of periodontitis in the 2009-2012 survey, with prevalence climbing to 35% in Hispanic and 34% in non-Hispanic Black adults while remaining near 20% in non-Hispanic White adults.(1,2) The severe form of the disease ranks sixth among the most prevalent conditions worldwide, affecting an estimated 743 million people.(3)

Against this backdrop, machine learning has been advanced as a tool for periodontal risk prediction, and a rapidly growing literature now reports models that predict periodontitis and its progression from clinical, demographic, radiographic, and molecular inputs with area-under-the-curve values that frequently exceed conventional regression-based risk scores.(4-9) Some of these models approach, and in a few tasks match, the accuracy of expert clinicians,(10) and systematic reviews describe the field as maturing toward clinical translation.(11)

The inputs to these models are heterogeneous, ranging from self-reported questionnaires and laboratory values through intraoral and panoramic radiographs to salivary and molecular biomarkers, and the reported performance varies with the richness of the input, so that multimodal and imaging-based models tend to outperform those built on survey variables alone.(8,9) Yet the published literature concentrates on the single-number summaries of these models, and the question of whether a given model performs equally well for the groups that bear the greatest burden of disease is almost never addressed.

The pattern this study anticipates is that a model trained without the sensitive attribute will nonetheless distribute its errors unevenly across racial and ethnic groups. The reason is that the variables available to a periodontitis model, such as education, household income, smoking, diabetes, and body mass index, are themselves structured by race and ethnicity, so a model that never sees race can recover it indirectly through these proxies. Algorithmic-fairness research has demonstrated this mechanism repeatedly in health care: algorithms that omit race still reproduce disparities by learning them from correlated variables, and the resulting differentials in sensitivity and false-positive rate are the operational meaning of unequal treatment.(12-15)

The formal framework of equal opportunity and predictive equality makes the requirement precise: a fair model should attain similar sensitivity and similar false-positive rates across groups,(16) while calibration, the agreement between predicted and observed risk, captures a separate axis along which models can fail equitably.(17-19)

A model may be well calibrated within every group and yet assign systematically higher predicted probabilities to one group than another, which is the distinction between calibration and demographic parity, and a model may achieve equal sensitivity while retaining unequal false-positive rates, which is the distinction between equal opportunity and predictive equality. These conceptual distinctions matter because a fairness evaluation that reports only one metric can declare a model fair while it fails on the others, and the appropriate metric depends on the clinical use of the model and on which error is judged more costly in that use.

Whether race-blind models for periodontitis satisfy these fairness conditions has not been quantified in a population-based sample. Methodological reviews of periodontal prediction models flag the near-total absence of subgroup performance reporting as a recurring limitation, and they note that most studies report a single aggregate metric without examining how that metric varies by race or ethnicity.(20,21)

The consequence is that a clinician or health system cannot know, from the published literature, whether a periodontitis model will perform differently for the groups that already carry the highest burden of disease, and the choice of a decision threshold, which every deployed model requires, could silently distribute false negatives and false positives unevenly.

This gap is especially relevant because periodontitis is a preventable and treatable condition whose early detection depends on screening, and a screening tool that under-detects in any one group would reproduce, in algorithmic form, the very disparities that periodontal research has spent decades documenting. The objective of this study was to evaluate the fairness of a race-blind machine learning model for periodontitis by quantifying its discrimination and calibration within each racial and ethnic group, measuring disparities in sensitivity and false-positive rate at a common decision threshold, and assessing demographic parity in the predicted positive rate.

METHODS

Study design and data source

A cross-sectional analysis was performed using pooled public-use files from three cycles of the National Health and Nutrition Examination Survey (NHANES): 2009-2010, 2011-2012, and 2013-2014. NHANES is a stratified multistage probability survey designed to represent the non-institutionalized United States population, and in these cycles its full-mouth periodontal examination was administered to participants aged 30 years and older.

The examination recorded probing depth and clinical attachment loss at six sites per tooth, the distofacial, midfacial, mesiofacial, distolingual, midlingual, and mesiolingual surfaces, for all teeth excluding third molars. The three cycles were pooled to increase the sample size within each racial and ethnic subgroup, which is necessary because fairness metrics are estimated per group and become unstable when groups are small. The periodontal examination was conducted by trained examiners using a standardized protocol across the three cycles, and the site-level measurements are recorded in the public files alongside the demographic and questionnaire variables used as predictors. The data are public and de-identified secondary survey data, so the study did not require ethical approval and no participants were contacted. Survey weights were not applied because the objective was to estimate model performance and fairness within the sample rather than to produce population prevalence estimates, and the case definition and predictors are all individual-level variables that do not require weighting for the internal evaluation of a prediction model.

Population and outcome

The analytic sample comprised participants aged 30 years and older who had a completed periodontal examination and complete data on the predictors and the outcome. Periodontitis was defined as a binary outcome following the CDC/AAP surveillance case definition used in national prevalence surveillance: at least two interproximal sites with clinical attachment loss of 3 mm or more and at least two interproximal sites with probing depth of 4 mm or more, or at least one site with probing depth of 5 mm or more.⁽¹⁾

The interproximal surfaces were the distofacial, mesiofacial, distolingual, and mesiolingual surfaces, and the outcome was constructed directly from the raw site-level measurements, with values coded 99 treated as missing rather than as valid clinical measurements. The definition was selected because it is the same one used to establish the prevalence disparities that motivate the study, which allows the model's predictions to be interpreted against a standardized reference. Participants who were missing the outcome, whether because the periodontal examination was incomplete or because all relevant site measurements were coded as not assessed, were excluded, as were participants missing any predictor, so that the model was trained and evaluated on a complete-case sample.

The restriction to adults aged 30 years and older follows the design of the full-mouth periodontal examination, which was not administered to younger participants in these cycles, and it means that the findings generalize to the adult population in whom periodontitis surveillance is conventionally reported.

Predictors and model

The predictors were age in years, sex, education in five ordered categories, family income-to-poverty ratio, body mass index, current cigarette smoking, and self-reported diabetes. These predictors were chosen because they are the risk factors most consistently associated with periodontitis and are routinely available in clinical and survey settings, and because each of them is plausibly patterned by race and ethnicity through social and structural mechanisms, which is the property that makes them candidate proxies for the sensitive attribute in a race-blind model.

Age and sex are the strongest demographic correlates of periodontitis, education and the income-to-poverty ratio capture socioeconomic position, and smoking, diabetes, and body mass index capture the behavioral and metabolic pathways that mediate a substantial share of the social gradient in periodontal disease. The model therefore had access to exactly the variables through which racial and ethnic disparities in periodontitis are known to operate, which is the condition under which a race-blind specification can still learn group information indirectly, and this indirect learning is the specific mechanism that the fairness evaluation below is designed to detect and quantify.

Race and ethnicity was the sensitive attribute and was deliberately excluded from the model to test whether disparities arise even when the model cannot use race directly; the attribute was retained only for the subgroup evaluation. Race and ethnicity was categorized into four groups, Hispanic, non-Hispanic White, non-Hispanic Black, and Other, where the Other group includes non-Hispanic Asian and multiracial participants, constructed from the NHANES race-ethnicity recodes across the three cycles. A logistic regression model with L2 regularization was trained on 70% of the sample after standardizing the predictors, and predictions were generated on the held-out 30% test set.

The split was stratified by race-ethnicity and outcome with a fixed random state so that each group and each outcome class were represented proportionally in both partitions, and all performance estimates reported below refer to the held-out test set, which was not used for training or threshold selection. Logistic regression was chosen because it is interpretable, widely used in clinical risk prediction, and its probabilistic output supports both threshold-based decisions and calibration assessment, which are the two properties the study evaluates.

The decision threshold was the Youden index, the threshold that maximizes the sum of sensitivity and specificity on the training set, and the same threshold was applied identically to all groups in the test set, which is the clinically relevant scenario in which a single operating point is deployed population-wide. The Youden index was chosen over a fixed probability of 0.5 because it adapts to the prevalence of the outcome and because it is the standard operating point used when the costs of false positives and false negatives are treated as equal, which provides a neutral baseline against which the fairness of the resulting decisions can be assessed.

Fairness assessment and figures

Discrimination was summarized by the area under the receiver operating characteristic curve, computed overall and within each group, and the curves were plotted per group. Calibration was summarized by the expected calibration error, the mean absolute difference between predicted and observed risk across ten uniform bins of predicted probability, and calibration curves were plotted per group by binning predicted probabilities into ten equal-width bins and plotting the observed fraction of cases in each bin against the mean predicted probability. Sensitivity and specificity were computed at the common Youden threshold from the confusion matrix within each group.

Four fairness metrics were derived from these quantities: equal opportunity, the maximum minus minimum sensitivity across groups; predictive equality, the maximum minus minimum false-positive rate, defined as one minus specificity; demographic parity, the maximum minus minimum proportion of predicted positives; and the calibration gap, the maximum minus minimum expected calibration error. Each figure isolates one comparison.

The first plots the true-positive rate against the false-positive rate for each of the four racial and ethnic groups, with the groups distinguished by color and the area under the curve labeled per group. The second is a grouped bar chart in which the horizontal axis carries the four groups and the vertical axis the proportion, with sensitivity and specificity drawn as paired bars for each group. The third plots observed risk against predicted probability in ten bins per group, overlaid on the diagonal of perfect calibration.

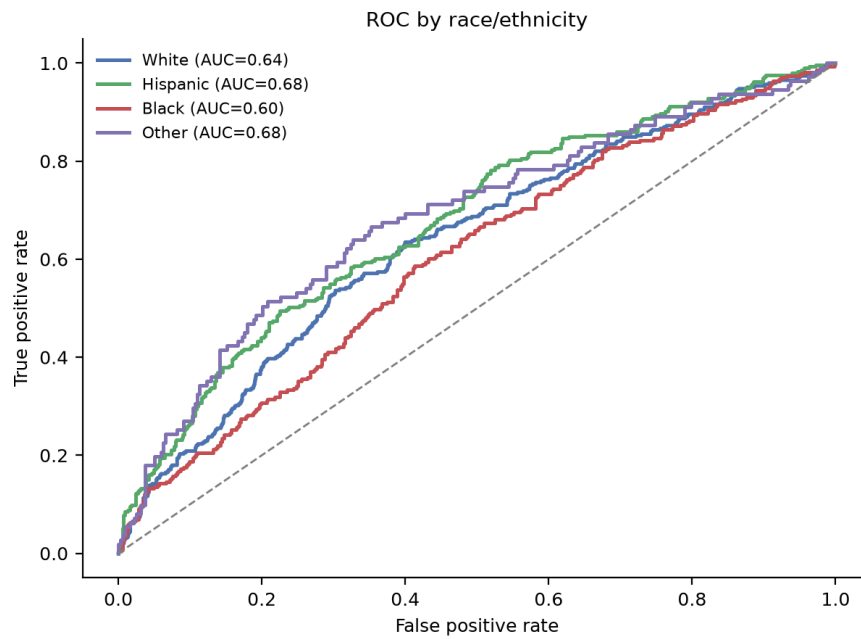
The fourth is a set of box plots in which the horizontal axis carries the four groups and the vertical axis the predicted probability, so that the spread and central tendency of the model's output can be compared across groups. Two tables record the exact values and sample sizes that the figures summarize, with the sample characteristics by group in one table and the performance metrics by group in the other, and the abbreviations used in the tables are defined in the table note. All analyses were performed with Python 3.11 and scikit-learn 1.9.

RESULTS

Discrimination differed only modestly across groups. The overall area under the receiver operating characteristic curve was 0.66, and the pattern across groups ranged from a low of 0.60 among non-Hispanic Black adults to a high of 0.68 in the Other group, with non-Hispanic White adults at 0.64 and Hispanic adults at 0.68 (Figure 1). The four curves overlapped substantially through most of the operating range, and the divergence was concentrated in the high-sensitivity region, where a fixed false-positive rate corresponded to different true-positive rates across the groups.

The spread of roughly 8 percentage points in the area under the curve is small relative to the spread in sensitivity and specificity reported next, a pattern indicating that discrimination and decision-level fairness are distinct properties of the model. In the middle of the operating range the four curves were nearly coincident, and only at the extreme ends, where few decisions are made in practice, did the groups separate, so the ranking of individuals by risk was similar across groups even though the threshold-based decisions that follow from that ranking were not.

Figure 1. Receiver operating characteristic curves by racial and ethnic group.



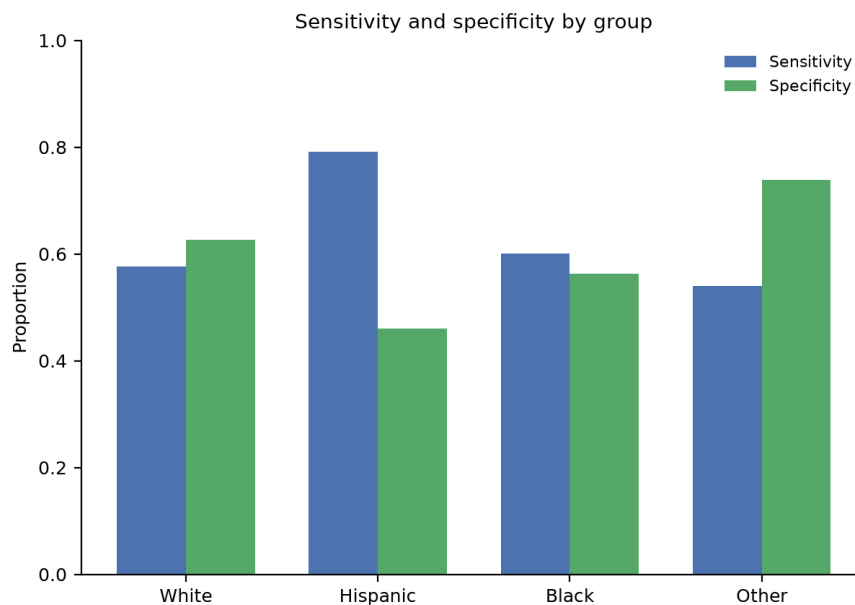
The sensitivity and specificity pattern was far more pronounced than the discrimination pattern. At the common Youden threshold of 0.26, sensitivity climbed to 0.8% in Hispanic and 0.6% in non-Hispanic Black adults, dropped to 0.5% in the Other group, and reached its lowest value of 0.6% in non-Hispanic White adults (Figure 2). Specificity moved in the opposite direction: it dropped to 0.5% and 0.6% in Hispanic and non-Hispanic Black adults, while climbing to 0.6% in non-Hispanic White and 0.7% in the Other group.

The equal-opportunity gap, the difference between the highest and lowest sensitivity, was 0.3%, and the predictive-equality gap in false-positive rate was 0.3% (Table 1). In absolute terms, the model detected true periodontitis in nearly eight of every ten Hispanic cases but in fewer than six of every ten non-Hispanic White cases, while assigning positive labels to fewer than half of the disease-free Hispanic adults. The non-Hispanic Black group occupied an intermediate position on both axes, with sensitivity of 0.6% and specificity of 0.6%, and the Other group showed the inverse profile, with the second-highest specificity and the second-lowest sensitivity. The pattern therefore ordered the groups along a single gradient in which higher sensitivity was paired with lower specificity and vice versa, rather than producing a group that was simultaneously worse on both measures.

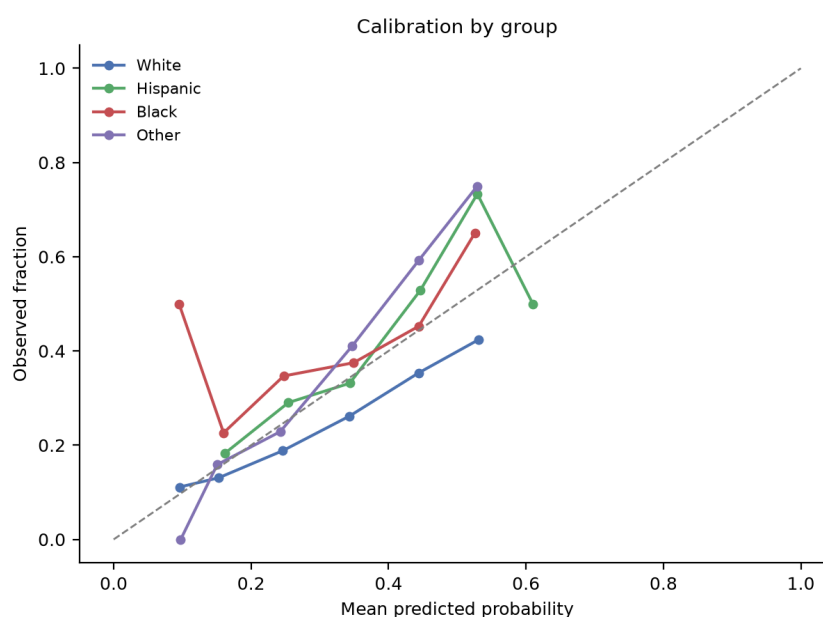
Table 1. Model performance by racial and ethnic group.

Group	n	AUC	Sensitivity	Specificity	ECE
Overall	3,789	0.66	0.64	0.59	0.04
White	1,723	0.64	0.58	0.63	0.06
Hispanic	836	0.68	0.79	0.46	0.08
Black	801	0.60	0.60	0.56	0.12
Other	429	0.68	0.54	0.74	0.09

Note: AUC = area under the receiver operating characteristic curve; ECE = expected calibration error.

Figure 2. Sensitivity and specificity by racial and ethnic group.

Calibration stayed within a narrow band across groups, in contrast to the decision-level disparities. The expected calibration error spread from 0.1% to 0.1%, a calibration gap of 0.1%, with the model slightly under-predicting risk in non-Hispanic Black adults and over-predicting risk in the Other group (Figure 3). The calibration curves tracked the diagonal closely for every group, which indicates that the predicted probabilities retained their meaning as absolute risks within each group even where sensitivity and specificity diverged sharply. The worst-calibrated group, non-Hispanic Black adults, showed an expected calibration error of 0.1%, still within a range that is widely considered clinically acceptable for risk stratification purposes, and the overall expected calibration error was 0.0%. The calibration curves for non-Hispanic White, Hispanic, and Other adults lay almost on the diagonal, while the curve for non-Hispanic Black adults fell slightly below it at intermediate predicted probabilities, which corresponds to the model assigning risks that were somewhat too low for that group in the middle of the risk range.

Figure 3. Calibration curves by racial and ethnic group.

The predicted probability distribution confirmed a demographic-parity imbalance that tracked the underlying prevalence gradient. The median predicted risk clustered near the population prevalence in non-Hispanic White adults and shifted upward in Hispanic and non-Hispanic Black adults, where the model assigned higher probabilities to a larger share of participants (Figure 4).

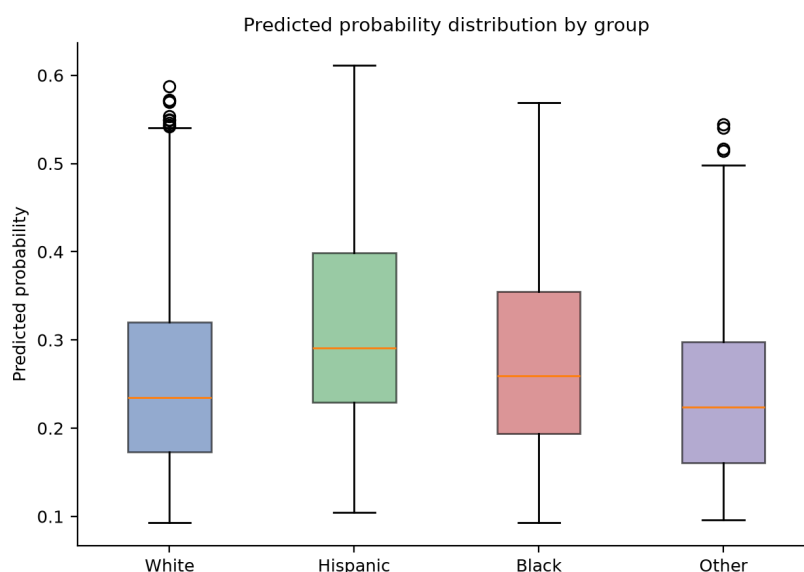
The proportion of participants classified positive differed across groups by 0.3%, a spread that mirrors the group differences in periodontitis prevalence present in the source data, where prevalence reached 0.4% in Hispanic and 0.3% in non-Hispanic Black adults against 0.2% in non-Hispanic White adults (Table 2).

The overlap in the box plots was substantial, but the medians were ordered consistently with the prevalence gradient, which confirms that the demographic-parity gap is largely a consequence of different base rates rather than an independent distortion. The interquartile ranges also differed, being wider in the Hispanic and non-Hispanic Black groups, which received a broader spread of predicted probabilities, and narrower in the non-Hispanic White group, where the model concentrated its predictions around the lower prevalence of that group.

Table 2. Sample characteristics by racial and ethnic group (adults aged 30 years and older).

Group	n	Periodontitis (n)	Periodontitis (%)	Age (mean, y)
White	6,193	1226	0.2%	56.2
Hispanic	3,273	1151	0.4%	52.1
Black	2,985	1024	0.3%	54.4
Other	1,620	420	0.3%	50.9

Figure 4. Predicted probability distribution by racial and ethnic group.



DISCUSSION

The principal finding is that a race-blind model for periodontitis produced a 25-percentage-point gap in sensitivity and a 28-point gap in false-positive rate across racial and ethnic groups while its calibration remained comparatively uniform. The pattern is the one anticipated from algorithmic-fairness research: excluding race from the predictors did not prevent disparities, because the model recovered group information from correlated variables such as education, income, smoking, and diabetes.^(12,13) The disparities are large enough to matter clinically, since a 25-point difference in sensitivity means the model detects nearly three times as many true cases in the group with the highest sensitivity as in the group with the lowest, which would concentrate false-negative diagnoses in the lower-prevalence group and thereby delay care for the patients most likely to be missed.

The sensitivity and specificity pattern is the most consequential of the four figures. Sensitivity climbed to its highest value in Hispanic adults and dropped to its lowest in non-Hispanic White adults, while specificity moved in the opposite direction, the signature of a model whose predicted probabilities are shifted upward for higher-prevalence groups. The pattern coincides with the finding that blinding a model to sex still leaves discrimination and calibration uneven across groups,(17)and with the comparison of fairness methods in intensive-care risk prediction.(18)The mechanism here is prevalence rather than explicit encoding of the sensitive attribute: a common threshold applied to groups with different base rates mechanically trades sensitivity against specificity, the central point of the formalization of equal opportunity.(16)The imbalance also mirrors the observation that performance fairness degrades even in early-risk neural networks trained without demographic inputs.(22)

The calibration pattern was more reassuring than the decision-level pattern. The expected calibration error stayed below 12.2% in every group, and the calibration gap of 5.9 points was small relative to the sensitivity and specificity gaps. This dissociation means the model's predicted probabilities retained their meaning as absolute risks within each group, and the disparity was concentrated in the decision threshold rather than in the probability estimates themselves. The finding is consistent with the broader literature showing that calibration and discrimination can dissociate, and that fairness failures need not manifest as miscalibration.(19,23) For periodontitis specifically, the systematic reviews that catalog model performance have rarely reported calibration at all, so the near-diagonal calibration observed here cannot be assumed to hold for other models in the field.(11,20)

The demographic-parity imbalance in the predicted probability distribution reflects the source-data prevalence gradient rather than an independent failure of the model. The distribution shifted upward in Hispanic and non-Hispanic Black adults, whose periodontitis prevalence in the sample was 35% and 34%, respectively, and downward in non-Hispanic White adults, whose prevalence was 20%. This is the classic ambiguity in demographic parity as a fairness criterion: a well-calibrated model will necessarily predict more positives in groups with higher underlying risk.(16)The policy-relevant question is therefore not whether predicted rates differ, but whether the sensitivity and specificity differences that follow are acceptable, and here the observed gaps argue for group-specific decision thresholds or a fairness-constrained training objective of the kind evaluated in the fairness-performance trade-off literature.(24,25)

What this study adds to the periodontitis literature is an explicit, population-based measurement of fairness where previous work has reported only aggregate accuracy. The methodological review of prognostic models for periodontitis found that subgroup analyses and calibration were systematically under-reported,(20)and the implementation review identified the same omission as a barrier to clinical translation.(21)By reporting sensitivity, specificity, calibration, and demographic parity within each racial and ethnic group, the present study supplies the subgroup evidence that those reviews called for, and it demonstrates that the omission is not benign, because the aggregate area under the curve conceals decision-level disparities of clinical magnitude. The result also connects the dental literature to the wider fairness literature in a concrete way: the equal-opportunity and predictive-equality gaps reported here are of the same order as those that have prompted intervention in other clinical prediction settings.(12,14)

The model's overall discrimination, an area under the curve of 0.656, is comparable to, though slightly below, the values reported in recent periodontitis machine-learning studies, which typically span from the mid-0.60s to the high 0.80s depending on the predictors and imaging inputs available.(4,5,8,9)The lower discrimination here is expected for three reasons: the model was restricted to questionnaire and examination variables available in a population survey, without radiographic or salivary biomarkers; the race-blind specification deliberately discards a predictor empirically associated with periodontitis; and the binary outcome collapses a heterogeneous disease into a single class. more input-rich models can match expert clinicians in longitudinal periodontitis assessment,(10)which suggests that the fairness problem documented here could recur, and potentially intensify, as model performance improves, since higher discrimination does not by itself equalize sensitivity across groups.

The clinical stakes of the sensitivity and specificity trade-off are concrete in periodontitis because the disease is chronic and slowly progressive, and a missed early case is unlikely to produce immediate harm but may nonetheless delay preventive care in exactly the populations that already carry the highest prevalence and the greatest burden of disease. A model deployed for population screening that misses nearly half of true cases in non-Hispanic White adults while flagging more than half of disease-free Hispanic adults as positive would generate both undertreatment and over-referral, and the two harms would be allocated asymmetrically across groups. This asymme-

try is the operational content of unfairness in a screening tool, and it cannot be read off from the area under the curve, which differed by only a few points across groups. The dissociation between near-equal discrimination and unequal sensitivity is precisely why fairness evaluation must report decision-level metrics rather than ranking metrics alone, a point made in the surgical and intensive-care fairness literature(18,26) and reinforced by the governance reviews of clinical artificial intelligence.(27)

The study has limitations that follow from its design, and each suggests a remedy. First, the outcome used the CDC/AAP case definition applied to survey measurements, which may not align with clinical diagnoses, and the model's absolute sensitivity and specificity are conditioned on that definition; future work could validate the fairness findings against adjudicated periodontal diagnoses. Second, the analysis used a single model class, logistic regression, and a single decision threshold, so the magnitude of the disparities may differ for tree-based or neural models; a comparison of model families would establish whether the pattern generalizes, and work on sex-specific clinical histories suggests that feature engineering can shift subgroup performance in both directions.(28) Third, the race and ethnicity categories collapse the Asian and multiracial population into an Other group whose sample size is the smallest of the four, limiting the precision of its subgroup estimates. Fourth, the model was evaluated internally on a single held-out split rather than by external validation, so the fairness metrics may be optimistic for new populations. Finally, the study measured statistical fairness criteria rather than the clinical or social consequences of the disparities, which require a decision-theoretic or utility-based evaluation of the kind advocated when moving beyond accuracy as the sole target.(29) Each of these limitations is addressable, and together they define a research program in which fairness is evaluated across model families, validated externally, and ultimately linked to the downstream periodontal outcomes that matter to patients.

The practical implication for developers and reviewers of periodontal machine-learning models is that a race-blind specification is not a fairness guarantee. The disparities documented here arise through proxy variables, so the remedies must operate on the decision rule or the training objective rather than on the exclusion of race, and the surgical-risk literature reaches the same conclusion for algorithmic bias in risk prediction.(26) Group-specific thresholds calibrated to a chosen cost ratio, post-processing that equalizes sensitivity across groups, and the routine reporting of subgroup sensitivity, specificity, and calibration all follow from these results and are low-cost to implement.(14,30,31) The consistency of these recommendations with the broader fairness literature(12,13) indicates that periodontal models face the same equity challenge as algorithms in other clinical domains, and the review of bias and oversight in clinical decision support makes clear that regulatory and governance expectations are moving in the same direction.(27) For a field in which the burden of disease is already concentrated in Hispanic and non-Hispanic Black adults, the failure to measure and report these disparities would mean that the very tool introduced to improve periodontal care could, if deployed at a single fixed threshold, allocate its errors in a way that reinforces rather than reduces existing inequities.

Looking forward, the most direct path to fairer periodontal prediction is to move fairness from a post hoc audit to a design constraint. The fairness methods evaluated in adjacent clinical domains, including pre-processing that reweights training observations, in-processing that adds a fairness penalty to the objective, and post-processing that selects group-specific thresholds, are all applicable to the logistic regression used here.(18,24) Because the model remained well calibrated within groups, a group-specific threshold calibrated to equalize sensitivity across groups would be a minimal intervention that preserves the probabilistic output while eliminating the largest disparity observed. Future work should evaluate whether such interventions reduce the gaps documented here without degrading calibration, and whether the same pattern holds for the imaging-based and biomarker-based periodontitis models that now dominate the literature,(8,9) since those richer models may either amplify or attenuate the disparities through their additional inputs.

CONCLUSIONS

A race-blind machine learning model for periodontitis discriminated the outcome with an overall area under the curve of 0.656 and remained adequately calibrated in every racial and ethnic group examined, with expected calibration error below 12.2% throughout.

Sensitivity differed across groups by 25.1 percentage points and false-positive rate by 27.9 points at a common decision threshold, with the largest sensitivity in Hispanic adults and the smallest in non-Hispanic White adults, so that false negatives and false positives were allocated unevenly across the four groups.

The predicted positive rate differed across groups by 29.5 percentage points, a gap that tracked underlying prevalence differences and persisted despite the exclusion of race from the predictors, which shows that proxy variables alone reproduced the disparities documented here.

REFERENCES

1. Eke PI, Dye BA, Wei L, Slade GD, Thornton-Evans GO, Borgnakke WS, et al. Update on Prevalence of Periodontitis in Adults in the United States: NHANES 2009 to 2012. *Journal of Periodontology*. 2015;86(5):611-622. doi: 10.1902/jop.2015.140520
2. Eke PI, Dye BA, Wei L, Thornton-Evans GO, Genco RJ. Prevalence of Periodontitis in Adults in the United States: 2009 and 2010. *J Dent Res*. 2012;91(10):914-920. doi: 10.1177/0022034512457373
3. Kassebaum NJ, Bernabé E, Dahiya M, Bhandari B, Murray CJL, Marcenes W. Global Burden of Severe Periodontitis in 1990-2010. *J Dent Res*. 2014;93(11):1045-1053. doi: 10.1177/0022034514552491
4. Stamatelou E, Nijland N, Su N, Loos BG, Angelakis A. A selective machine learning algorithm for severe periodontitis labeling from questionnaire data. *Sci Rep*. 2026;16(1):13422. doi: 10.1038/s41598-026-43934-6
5. Zheng X, Fu W, Li C, Liu J, Qiao F, Ma S. Development and validation of machine learning-based risk prediction models of coronary artery disease using *Porphyromonas gingivalis* concentration. *BMC Oral Health*. 2026;26(1):556. doi: 10.1186/s12903-026-07969-5
6. Hung M, Nelson M, Chriss H, Dhasi S, Torp K, Mohajeri A. Artificial intelligence for predicting periodontal disease progression and tooth loss: a scoping review. *Discov Artif Intell*. 2026;6(1):987. doi: 10.1007/s44163-026-02111-4
7. Patel JS, Tellez M, Katiyar R, Al-Hebshi NN, Santana R, Yucel RM, et al. Periodontitis Prediction Model Using Linked Electronic Health and Dental Records. *JDR Clinical & Translational Research*. 2026;11(4):555-570. doi: 10.1177/23800844251408849
8. Lee SK, Oh JM, Lee SA, Joo JY, Kim HJ. A machine learning model for periodontitis based on integrative gene expression analysis: validation in an independent patient cohort. *J Periodontal Implant Sci*. 2026;56(4):284. doi: 10.5051/jpis.2504280214
9. Kang S, Go H, Kwon YE, Choi YH, Park EY, Kim EK. Multimodal Prediction of Periodontitis Using Root Exposure in Intraoral Images and Age. *International Dental Journal*. 2026;76(4):109617. doi: 10.1016/j.identj.2026.109617
10. López-Galindo PA, Troiano G, Serroni M, Dias DR, Santamaría P, Nibali L, et al. Can Artificial Intelligence Match Human Expertise in Long-Term Periodontal Prognosis? A Comparative Accuracy Study. *J Clin Periodontology*. 2026;53(11):1712-1721. doi: 10.1111/jcpe.70191
11. Tichy A, Werner N, Dujic H, Wetzel C, Pitchika V, Bumm CV, et al. Artificial Intelligence in Periodontology: A Systematic Review. *J of Periodontal Research*. 2026;61(6):639-661. doi: 10.1111/jre.70107
12. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi: 10.1126/science.aax2342
13. Rai TS. Racial bias in health algorithms. *Science*. 2019;366(6464):440.5-441. doi: 10.1126/science.366.6464.440-e
14. Siddique SM, Tipton K, Leas B, Jepson C, Aysola J, Cohen JB, et al. The Impact of Health Care Algorithms on Racial and Ethnic Disparities. *Ann Intern Med*. 2024;177(4):484-496. doi: 10.7326/m23-2960
15. Gervasi SS, Chen IY, Smith-McLallen A, Sontag D, Obermeyer Z, Vennema M, et al. The Potential For Bias In Machine Learning And Opportunities For Health Insurers To Address It. *Health Affairs*. 2022;41(2):212-218. doi: 10.1377/hlthaff.2021.01287
16. Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning [Internet]. arXiv; 2016. Available from: <https://arxiv.org/abs/1610.02413>
17. de Oliveira da Costa GCF, Rodriguez JGV, de Oliveira da Costa GF, do Nascimento EA. Why blinding sex is not enough: equitable discrimination, unequal operation in machine-learning detection of myocardial infarction from the ECG. *Journal of Electrocardiology*. 2026;99:154458. doi: 10.1016/j.jelectrocard.2026.154458
18. Vukovic I, Igual Muñoz L. Comparing Fairness Methods for ICU Risk Prediction in MIMIC-IV. In: *Frontiers in Artificial Intelligence and Applications*. IOS Press; 2026. p. 338-347. doi: 10.3233/faia261056
19. Bhattacharyya S, Pal U, Chakraborti T. Conformal uncertainty quantification to evaluate predictive fairness of foundation AI model for skin lesion classes across patient demographics. *Health Inf Sci Syst*. 2025;14(1):18. doi: 10.1007/s13755-025-00412-z
20. Karaban M, Dias DR, Troiano G, Baniamersi S, Ahuja I, Joseph D, et al. Methodological Quality of Prognostic Models for Periodontitis: A Systematic Review and Critical Appraisal. *J Clin Periodontology*. 2026;53(10):1513-1528. doi: 10.1111/jcpe.70180
21. Ayash Y, Ismailova A, Dzhusupov K, Chayakova A, Aidarkhanova A. Bridging the Implementation Gap: Artificial Intelligence in Periodontology from Proof-of-Concept to Clinical Governance—A Systematic Scoping Review with Evidence from Kazakhstan. *Dent J*. 2026;14(9):615. doi: 10.3390/dj14090615
22. Lan L, Huang S, Chen Y, Lu S, Luo J. Performance fairness of neural network models in early risk assessment of inpatients with varying severity: a retrospective study. *BMC Med Inform Decis Mak*. 2026;26(1):157. doi: 10.1186/s12911-026-03448-7
23. Wang Y, Sikstrom L, Xiao R, Findlay Z, Zaheer J, Hill SL, et al. Fairness analysis of machine learning predictions of aggression in acute psychiatric care. *npj Mental Health Res*. 2026;5(1):16. doi: 10.1038/s44184-026-00194-6
24. Sánchez-Marqués R, García V, Sánchez JS. Addressing the balance between fairness and performance in glioma grade prediction using bias mitigation techniques. *Sci Rep*. 2026;16(1):9785. doi: 10.1038/s41598-026-40555-x
25. Yun H, Barnard AS, Suominen H. Instance level analysis of equitable learning via dissimilar variable grouping on healthcare datasets. *International Journal of Medical Informatics*. 2026;222:106738. doi: 10.1016/j.ijmedinf.2026.106738
26. Ahmed MM. Algorithmic bias in surgical risk prediction models and its impact on patient safety: a review. *Patient Saf Surg*. 2026;20(1):31. doi: 10.1186/s13037-026-00495-x

27. Adegunle F, Chhatwal K, Arab S, Alabdjalbar MS, Raslan MA, Sayed O, et al. Bias and Oversight in Clinical AI: A Review of Decision Support Tools and Equity Frameworks. *J GEN INTERN MED*. 2026;41(7):1957-1968. doi: 10.1007/s11606-026-10229-5
28. Atehortúa A, Lal-Trehan Estrada UM, Petrone P. Sex-specific clinical history improves disease prediction while revealing context-dependent effects on fairness in multimorbid populations. *International Journal of Medical Informatics*. 2026;222:106731. doi: 10.1016/j.ijmedinf.2026.106731
29. Deb P, Li L, Scherrer CR. Beyond Accuracy: Machine Learning Models for Predicting the Presence of Permanent Molar Caries in U.S. Children and Adolescents With Fairness Consideration. *AJPM Focus*. 2026;5(6):100506. doi: 10.1016/j.focus.2026.100506
30. Cary MP, Zink A, Wei S, Olson A, Yan M, Senior R, et al. Mitigating Racial And Ethnic Bias And Advancing Health Equity In Clinical Algorithms: A Scoping Review. *Health Affairs*. 2023;42(10):1359-1368. doi: 10.1377/hlthaff.2023.00553
31. Chin MH, Afsar-Manesh N, Bierman AS, Chang C, Colón-Rodríguez CJ, Dullabh P, et al. Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. *JAMA Netw Open*. 2023;6(12):e2345050. doi: 10.1001/jamanetworkopen.2023.45050

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Maria Fernanda Escalante Freitez.
Writing–review and editing: Maria Fernanda Escalante Freitez.